Exercise 8

1. In this exercise we are investigating the ozone dataset which you have already seen in the lecture. The dataset `ozone` is available in numerous R-packages, e.g. in the package `gss`. You can load it with `data(ozone, package = "gss")`. If you do not have access to the required R-package you can get it via
`read.table("http://stat.ethz.ch/Teaching/Datasets/ozone.dat", header = TRUE)`.

A short description of the variables is available at `help(ozone, package = "gss")` or at `http://stat.ethz.ch/Teaching/Datasets/ozone.txt`.

a) Get an overview of the data with `pairs()`. You should take the log of the response (`upo3`) and remove the outlier in the predictor `wdsp`.

b) We want to use MARS as regression method. To decide on the "best" interaction degree we use 10-fold cross validation (see 4.3.2). As there are 329 data points you can split the dataset into 9 sets consisting of 33 points and an additional set of 32 points. You should use the sum of squared errors as loss function. Try interaction degrees 1, 2, 3 and 4. Which one performs best? Check the model-assumptions (Tukey-Anscombe plot) for the selected model.

R-Hints:
`mars()` is available in the package `mda`.
Usage: `mars(x, y, degree = 2)`, where `x` is the design matrix containing the predictors, `y` is a vector containing the response variable and `degree` is the maximum interaction degree. For more information see `help(mars)`.

c) We now want to visualize the effects of the predictors. Therefore we restrict ourselves to a model with interaction degree equal 2. If your model from b) is different, change it for this exercise to interaction degree 2.
If `fit.deg2` is the MARS fit with interaction degree 2,
`unique(abs(fit.deg2$factor[fit.deg2$selected.terms]))` will give you a matrix. The rows are the terms in the model, the columns are the corresponding predictors: If there is a 1 at position (i, j) it means that the i^{th} term contains the j^{th} predictor. For the example below it follows that term 1 is the intercept, term 2 is the main effect of `sbtp` and term 3 is the interaction between `hmdt` and `ibht`.

```
     vdht wdsp hmdt sbtp ibht dgpg ibtp vsty day
[1,]    0    0    0    0    0    0    0    0   0
[2,]    0    0    0    1    0    0    0    0   0
[3,]    0    0    1    0    1    0    0    0   0
```

To visualize the j^{th} main effect you can create a design matrix where all columns are constant except for the j^{th}.
For the interactions you should set up a grid and set all other predictors constant as above.
The following R-code may give you an idea on how to proceed.

R-Hints:

```
model.m <- unique(abs(fit.deg2$factor[fit.deg2$selected.terms,]))
nrvars <- apply(model.m, 1, sum)
ind.main <- which(nrvars == 1) ## main effects
ind.int <- which(nrvars == 2)  ## interactions

## Design matrix consisting of a replication of the predictor means
Xp <- matrix(apply(d.ozone.e[,-1], 2, mean), nrow = nrow(d.ozone.e),
             ncol = ncol(d.ozone.e[-1]), byrow = TRUE)
## Range for each predictor
xr <- apply(d.ozone.e[,-1], 2, range)

## Visualize the main effects
par(mfrow = c(4,2))
for(i in ind.main){
  j <- which(model.m[i,] == 1) ## predictor
  Xp1 <- Xp
  Xp1[,j] <- seq(???, ???, ???)
  Xf <- predict(???)
  plot(Xp1[,j], Xf, ???)
}

## Visualize the interactions

dens <- 20 ## density of the grid
Xp <- matrix(apply(d.ozone.e[,-1], 2, mean), nrow = dens^2,
             ncol = ncol(d.ozone.e[-1]), byrow = TRUE)

par(mfrow = c(3,1))
for(i in ind.int){
  j.m <- which(model.m[i,] == 1) ## 2 predictors
  v1 <- seq(xr[1,j.m[1]], xr[2, j.m[1]], length.out = dens)
  v2 <- seq(xr[1,j.m[2]], xr[2, j.m[2]], length.out = dens)
  grid.p <- as.matrix(expand.grid(v1, v2)) ## create grid points
  Xp1 <- Xp
  Xp1[,j.m] <- grid.p
  Xf <- predict(???, ???)
  persp(v1, v2, matrix(Xf, nrow = length(v1)), ltheta = ???,
        shade = 0.75, ???)
}
```

Preliminary discussion: Friday, June 3, 2005. **Deadline:** Friday, June 10, 2005.

Advice: Thursdays from 12.00-13.00, LEO C12.1, Leonhardstr. 27. Or contact either Maik Berchtold, `berchtld@stat.math.ethz.ch`, or Lukas Meier, `meier@stat.math.ethz.ch`.